Relation Mining Experiments in the Pharmacogenomics Domain

Rinaldi, Fabio ; Schneider, Gerold ; Clematide, Simon

Abstract: The mutual interactions among genes, diseases, and drugs are at the heart of biomedical research, and are especially important for the pharmacological industry. The recent trend towards personalized medicine makes it increasingly relevant to be able to tailor drugs to specific genetic makeups. The pharmacogenetics and pharmacogenomics knowledge base (PharmGKB) aims at capturing relevant information about such interactions from several sources, including curation of the biomedical literature. Advanced text mining tools which can support the process of manual curation are increasingly necessary in order to cope with the deluge of new published results. However, effective evaluation of those tools requires the availability of manually curated data as gold standard. In this paper we discuss how the existing PharmGKB database can be used for such an evaluation task in a way similar to the usage of gold standard data derived from protein-protein interaction databases in one of the recent BioCreative shared tasks. Additionally, we present our own considerations and results on the feasibility and difficulty of such a task.

DOI: https://doi.org/10.1016/j.jbi.2012.04.014

Relation Mining Experiments in the Pharmacogenomics Domain

Fabio Rinaldi[a,*], Gerold Schneider[a], Simon Clematide[a]

[a] *Institute of Computational Linguistics, University of Zurich, Binzmühlestrasse 14, 8050 Zürich, Switzerland*

Abstract

The mutual interactions among genes, diseases, and drugs are at the heart of biomedical research, and are especially important for the pharmacological industry. The recent trend towards personalized medicine makes it increasingly relevant to be able to tailor drugs to specific genetic makeups. The pharmacogenetics and pharmacogenomics knowledge base (PharmGKB) aims at capturing relevant information about such interactions from several sources, including curation of the biomedical literature.

Advanced text mining tools which can support the process of manual curation are increasingly necessary in order to cope with the deluge of new published results. However, effective evaluation of those tools requires the availability of manually curated data as gold standard.

In this paper we discuss how the existing PharmGKB database can be used for such an evaluation task in a way similar to the usage of gold standard data derived from protein-protein interaction databases in one of the recent

* *Corresponding author*:
Tel: +41 44 635 7132; Fax: +41 44 635 6809;
Email addresses: fabio.rinaldi@uzh.ch (Fabio Rinaldi), gschneid@cl.uzh.ch (Gerold Schneider), simon.clematide@uzh.ch (Simon Clematide)

BioCreative shared tasks. Additionally, we present our own considerations and results on the feasibility and difficulty of such a task.
Keywords:
text mining, pharmacogenomics, literature curation

---

1. Introduction

In recent years the management of the vast amount of knowledge generated by the intensive growth of the biomedical sciences has been recognized as a scientific problem in its own right. Various approaches have been proposed in order to organize this very broad knowledge space through a variety of techniques. Ontologies, controlled vocabularies, and curated databases, are some of the instruments that are being developed in order to help scientists and practitioners to organize and easily access the vast amount of information that is typical of this domain.

Although there is an increasing tendency towards direct submission of experimental data and results to reference repositories, it is still the case that most of the produced knowledge is only available in the format of scientific publications. Such knowledge is most commonly extracted through the intellectually intensive and time consuming process known as literature curation, whereby highly-trained domain experts are employed in order to read the publications and distill from them the relevant information for a particular biomedical task. Since this approach cannot possibly keep up with the very intensive rate at which new results are being published [1], it is helpful to consider the usage of text mining tools, derived from research in natural language processing, which allow a partial automation of this task, and can be

used as supporting tools for human curators. Novel tools have the potential for enhancing the performance of database curators, speeding up their work and increasing their productivity by performing automatically their most tedious functions and allowing them to focus their valuable skills on the most rewarding functions of their activity.

Several text mining approaches have been described in the literature. However these publications seldom allow a comparative evaluation of the performance of the systems, due to the broad nature of tasks and corpora which are the subjects of investigation. In order to allow for a verifiable comparison under controlled circumstances, the text mining community has recently engaged in a number of comparative evaluations called "shared tasks", which are run in the format of a competition. The organizers of such shared tasks are responsible for delivering annotated training data and unannotated test data to the participants, and for scoring the results of the participating system on the test data, using a set of reliable metrics. The participants tune their systems using the provided training data and then have a limited amount of time to run their systems over the test data and deliver their results back to the organizers. Some of the best-known recent shared tasks are BioCreative [2, 3], the BioNLP shared task [4], and CALBC [5].

Each of these competitive evaluations typically involves several independent tasks, such as the recognition of mentions of specific domain entities in text, their normalization to specific database identifiers, and recognition of interactions among entities. For example, BioCreative includes a Gene Normalization task which involves detection of mentions of genes and their disambiguation to EntrezGene identifiers. An example of a relation mining

task, performed again in BioCreative, is the detection of mentions of protein-protein interactions. Another example of relation mining is the detection of specific event types (e.g. 'regulation', 'binding') and their arguments in the BioNLP shared task.

While the general philosophy of these shared tasks is similar, they differ substantially in the type of data that they use as a reference for the definition of the tasks. The data used by BioCreative is sourced from existing and widely known databases (e.g. IntAct [6], MINT [7], BioGrid [8]) and adapted to the specific requirements of the BioCreative tasks. In agreement with the database providers some of their curated data is temporarily retained from publication in order to be used for testing. The BioNLP evaluation uses as data their own annotated corpora, produced within the scope of the GENIA project [9], which allows more complex tasks, thanks to the better structured in-text annotation. CALBC targets the harmonization of entity annotations across several text mining tools, and therefore does not need a reference corpus, but rather aims at building a large 'consensus' corpus without recurring to manual annotation and verification.

The approach used in the BioCreative shared task is therefore more directly relevant for database curation teams, as it uses data derived from existing databases, and additionally aims at directly supporting the process of curation by stimulating the incremental improvements of tools directly relevant for specific subtasks. BioCreative has had so far three main editions (2003, 2006, 2010) and an intermediate edition (2009).[1] The OntoGene

[1]The organization of the competition has involved several groups, including Lynette Hirschman (MITRE, US), Alfonso Valencia and his group (CNIO, Spain), Cathy Wu and

group participated in the protein-protein interaction task of 2006, obtaining competitive results [10], and 2009, obtaining the best published results [11]. Additionally, they participated (with highly ranked results) in the 'interaction method' task of 2006 [12] and in all tasks of the 2011 evaluation [13, 14, 15, 16].

In the interaction task the participants, starting from the unannotated raw text of the journal papers, have to identify protein-protein interactions mentioned in the documents. In the evaluation phase these results will be compared with interactions previously identified by expert curators. The task is very challenging as it requires the identification of mentions of relevant proteins, their disambiguation using database identifiers (UniProt) and the identification of mentions of potential interactions. Furthermore, not all interactions mentioned in the paper are considered, but solely those that are reported by the authors as their main research results. Our official results [11] were the best reported according to the official metric, the so called 'raw AUC iP/R' , which provides a good indication of the quality of the ranking of the results [17].

One of the problems that organizers of a text mining shared task need to face is the availability of suitable primary annotations. While there are numerous databases that curate protein-protein interactions, availability of annotations for other types of interaction is scarce. The PharmGKB database [18, 19] curates interactions among drugs, diseases and genes, with a specific focus on genetic polymorphism and its relationships to disease susceptibility

Cecilia Arighi (U. Delaware, US), Kevin Cohen (U. Colorado, US), W. John Wilbur and his team (NLM, US).

and drug response. The availability of the curated interactions for download renders the PharmGKB an interesting resource for the development and testing of text mining systems.

In this paper we describe how the PharmGKB database can be used as a "gold standard" in a text mining task analogous to the protein-protein interaction task as practiced in the BioCreative competitive evaluations. We show how the available data can be straightforwardly converted into a suitable format, and how the same tools used for scoring BioCreative results can be applied to this dataset. We then describe our own approach aimed at mining such interactions using the OntoGene text mining system and we present results recently obtained. While not yet optimal, such results can certainly serve as a baseline reference for further developments in this area. Finally we present our interactive curation system ODIN and the specific adaptation to the PharmGKB dataset.

2. Methods and Results

In this section we first characterize the resources that we have used for the experiments described in this paper, then propose evaluation methods derived from the experience of BioCreative. Next we describe in detail our basic relation mining approach, followed by a syntax-based enhancement aimed at high-precision retrieval. Finally we describe how the ranking of the results can be optimized using a machine-learning approach and discuss our results.

2.1. Resources

The PharmGKB database provides a very rich collection of manually curated resources concerning how human genetic variation leads to differing responses to drugs. Access to this data is provided through sophisticated web interfaces. Additionally, PharmGKB offers free download of their data in simple textual formats (tab-separated values). The resources available for download include lists of all domain-relevant entities (genes, diseases, drugs) used by PharmGKB curators, and a list of all interactions annotated by them.

Each conceptual entity, uniquely identified by a PharmGKB identifier (ID), comes with additional information such as a set of terms which could be used by authors in scientific publications to refer to it, as well as additional identifiers that allow its mapping into other reference databases: EntrezGene, Emsembl and UniProt for genes, MeSH, SnoMedCT, UMLS for diseases, and ATC for drugs. Relationships are represented as binary interactions between two typed IDs (a standard name is also provided for each entity), with supporting evidence provided in form of the PubMed IDs of the publications which mention the specific interaction.

```
1. Drug:PA450428 methotrexate Disease:PA165817398 Myelosuppression   PMID:17323057,PMID:20335220
2. Gene:PA238    MAPT         Disease:PA446850    Blindness,Cortical PMID:9804125
```

For example, the previous two lines from the relationship file describe two specific interactions between (1) the drug *methotrexate* and the disease *Myelosuppression*, (2) the gene *MAPT* and the disease *Cortical Blindness*. Notice that in this format evidence can come from multiple publications. For

a number of relationships involving genetic polymorphisms, an additional reference to the Single Nucleotide Polymorphism database at NCBI (dbSNP)[2] is provided in the form of a rsID (reference single-nucleotide polymorphism [SNP] ID). Interactions that are recognized as playing an important role in a known pathway are additionally annotated with a reference to the specific pathway (which is described in a separate file).

There are a total of 22827 interactions available in the version of PharmGKB which we have used for the experiments described in this paper.[3] Once the multiple evidence sources for each interaction are separated, we obtain a total of 36557 triples consisting of two entity IDs and one source IDs. These triples can be classified according to the type of the source, giving the following distribution:

```
26122 PMID
 5467 Pathway
 4968 rsID
```

In our experiments we consider only the interactions which are supported by a PubMed identifier, discarding the pathway-based and rsID-based interactions. These 26122 binary interactions, which are based upon 5062 distinct articles,[4] can be used as a "gold standard" in a text mining task analogous to the BioCreative protein-protein interaction task [2, 3].

[2] `http://www.ncbi.nlm.nih.gov/projects/SNP/`

[3] All numerical data provided in this paper refer to a version of PharmGKB downloaded in September 2010.

[4] In our experiments we could effectively access only 5045 articles, for a total of 24278 non-reflexive relations.

For our experiments, we decided to use only the entities provided by PharmGKB itself (drugs, genes, diseases), which are distributed as follows:

```
Drugs: 30351 terms / 2986 ids
Diseases: 28633 terms / 3198 ids
Genes:  176366 terms / 28633 ids
```

If novel, unseen articles have to be processed, these terminologies can be easily extended using the same databases used by PharmGKB (e.g. Entrez-Gene for genes). However any new entity would not yet have a PharmGKB identifier, so it would be impossible to use it in a validation task such as the one that we are discussing.

The interactions in the PharmGKB 'gold standard' can be classified according to the types of the interacting entities, leading to the following distribution (directionality of the interaction is ignored):

```
 10597  Gene/Drug
  9415  Gene/Disease
  4202  Drug/Disease
   928  Gene/Gene
   742  Drug/Drug
   238  Disease/Disease
```

2.2. Evaluation Methods

The format of the relationship file provided by PharmGKB lends itself to easy transformation into a format equivalent to the one used for the protein-protein interaction task of BioCreative II.5 [3]. Given a text mining tool

which can produce a ranked list of gene/drug/disease interactions, it becomes possible to score these results against the PharmGKB-derived data using a scoring tool provided by the BioCreative organizers.

The BioCreative scorer returns an evaluation of the results according to the standard metrics used in information retrieval (Precision, Recall, F-score) as well as a relatively novel measure called "AUC iP/R" (area under the curve of the interpolated precision/recall graph).[5] The purpose of the AUC iP/R measure (henceforth "AUC", not to be confused with the more frequently used "AUC of the ROC curve" metric) is to provide an indication of the quality of the ranking of the results. The intuitive idea is that, given equivalent P/R/F figures, correct predictions which occur towards the top of the ranked list of results are more useful than results which are lower in the ranking. The implicit assumption is that a curator could use the ranking to decide where to stop looking at the results, therefore a better ranking provides a better user experience.

All numerical results are provided in 'micro' and 'macro' mode. Micro means that all interactions from all articles are pooled together and evaluated as one block. Macro means that results are computed on each article, and then averaged. For the macro results, standard deviations are also provided. Figure 1 shows two examples of the full results returned by the BioCreative scorer. The micro average numbers do not reflect the mean per-document quality if a lot of documents contain only one relevant relation, and a few

[5]The AUC iP/R curve is defined in [20], a detailed operative description of AUC iP/R, as used in the BioCreative evaluations, can be found at `http://www.biocreative.org/tasks/biocreative-ii5/biocreative-ii5-evaluation/`

```
Evaluated documents:       4870
Evaluated results:       375095

Hits TP: 9249 FP: 365846 FN: 14257

Global test-set results (micro-averaged)
Micro precs.: 0.02466 recall: 0.39347
f-scr.: 0.04641
Micro AUC iP/R: 0.06512

Average per-document results (macro-averaged)
StdDev precs.: 0.13941 recall: 0.43879
f-scr.: 0.15110
StdDev AUC iP/R: 0.36454
Macro precs.: 0.06707 recall: 0.60624
f-scr.: 0.09681
Macro AUC iP/R: 0.33797
```

```
Evaluated documents:       4831
Evaluated results:       161820

Hits TP: 6096 FP: 155724 FN: 17213

Global test-set results (micro-averaged)
Micro precs.: 0.03767 recall: 0.26153
f-scr.: 0.06586
Micro AUC iP/R: 0.05290

Average per-document results (macro-averaged)
StdDev precs.: 0.16205 recall: 0.44470
f-scr.: 0.17530
StdDev AUC iP/R: 0.36987
Macro precs.: 0.09030 recall: 0.48380
f-scr.: 0.12302
Macro AUC iP/R: 0.31251
```

Figure 1: Initial results, as reported by the BioCreative scoring utility, obtained through pairwise combination of all entities detected in the whole abstract (left) or sentence-by-sentence (right).

documents contain many relevant relations. This is the case for PharmGKB, where certain documents contain hundreds of relations and almost 2000 documents contain only one relation, as can be seen in Figure 2. 40% of the documents contain just one relation. However these 40% of documents contribute less than 10% of all relations. Approx. 90% of the documents contain 10 or fewer relations. However these documents contain less than 50% percent of all relations.

If *all* candidates generated by a system are considered for evaluation, then P/R/F are not influenced by their ranking. The only measure which is influenced by it is the "AUC iP/R" metric. Intuitively, a higher value of

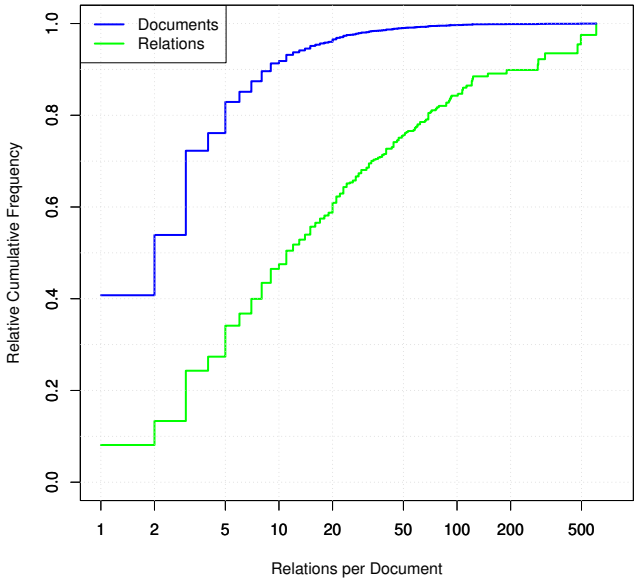


Figure 2: Relative cumulative frequency of PharmGKB documents and the number of relation occurrences contained in each document. The x axis shows the number n of relations in a document. The blue line plots the percentage of documents containing at most n relations. The green line shows the percentage of relations which are found in documents containing at most n relations.

AUC means that correct results tend to appear higher up in the ranking. Given the same set of results, a better ranking implies that any given cut-off threshold will result in higher precision (and lower recall), while an optimal threshold can easily be selected to maximize the F-score. The "AUC iP/R" metric has been criticized for being biased towards recall. It is in fact possible to improve the AUC by simply adding more results (provided at least some of them are correct) to the set of delivered results, even if this might cause a

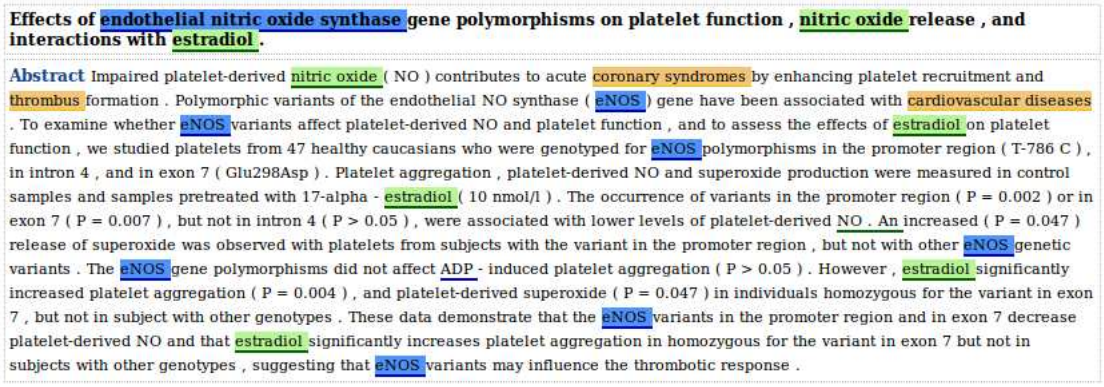
significant drop in precision.

A recently proposed alternative measure of the ranking of the results is the "Threshold Average Precision" (TAP-k) [21], which (in slightly simplified terms) averages precision for the results above a given error threshold. While the TAP-k metric is easier to interpret and directly relevant for the end user, who in most cases would not be willing to inspect a long list of results containing many false positives, we remain convinced that the AUC score offers a better way to directly compare fully automated text mining systems *over their entire retrieval spectrum*. In other words, comparing AUC values obtained using a threshold or filtering of the results is not particularly meaningful, as the loss of recall will also have an impact on AUC. Therefore we suggest to always use together both TAP-k (using a small fixed set of k values) and AUC at maximal recall.

2.3. Interaction Mining experiments

For our experiments, we automatically download from PubMed (using the `efetch` script from Entrez utilities)[6] the abstracts corresponding to the PubMed IDs mentioned by the PharmGKB relationship file. All experiments described in this paper are based on this collection of abstracts. It would of course be desirable to work on full papers rather than abstracts, however not all these publications are freely downloadable, and most importantly, they are not available in a common format. The lack of a common format hinders the usability of full-text publications, as it makes it more difficult to identify significant zones of the papers (e.g. results sections) or zones that require

[6] `http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html`



Effects of endothelial nitric oxide synthase gene polymorphisms on platelet function , nitric oxide release , and interactions with estradiol .

Abstract Impaired platelet-derived nitric oxide (NO) contributes to acute coronary syndromes by enhancing platelet recruitment and thrombus formation . Polymorphic variants of the endothelial NO synthase (eNOS) gene have been associated with cardiovascular diseases . To examine whether eNOS variants affect platelet-derived NO and platelet function , and to assess the effects of estradiol on platelet function , we studied platelets from 47 healthy caucasians who were genotyped for eNOS polymorphisms in the promoter region (T-786 C) , in intron 4 , and in exon 7 (Glu298Asp) . Platelet aggregation , platelet-derived NO and superoxide production were measured in control samples and samples pretreated with 17-alpha - estradiol (10 nmol/l) . The occurrence of variants in the promoter region (P = 0.002) or in exon 7 (P = 0.007) , but not in intron 4 (P > 0.05) , were associated with lower levels of platelet-derived NO . An increased (P = 0.047) release of superoxide was observed with platelets from subjects with the variant in the promoter region , but not with other eNOS genetic variants . The eNOS gene polymorphisms did not affect ADP - induced platelet aggregation (P > 0.05) . However , estradiol significantly increased platelet aggregation (P = 0.004) , and platelet-derived superoxide (P = 0.047) in individuals homozygous for the variant in exon 7 , but not in subject with other genotypes . These data demonstrate that the eNOS variants in the promoter region and in exon 7 decrease platelet-derived NO and that estradiol significantly increases platelet aggregation in homozygous for the variant in exon 7 but not in subjects with other genotypes , suggesting that eNOS variants may influence the thrombotic response .

Figure 3: Annotated abstract: genes are highlighted in blue, diseases in yellow, drugs in green.

special processing (e.g. tables).

Our main aim is to show that the PharmGKB dataset represents an interesting resource for the evaluation of text mining tools, in particular in relation to the detection of binary interactions other than the already widely studied protein-protein interactions. In this respect, we regard our experiments with abstracts as a relevant proof of concept, even if we intend to consider the full text of the PharmGKB papers in future work.

We apply our OntoGene relation mining system (OG-RM, [22, 11]) in order to annotate the input documents, using only the terminology provided by PharmGKB (see example in Figure 3). First, in a preprocessing stage, the input text is transformed into a custom XML format, and sentences and token boundaries are identified. For these tasks, the LingPipe[7] tokenizer and sentence splitter, which have been trained on biomedical corpora, are used.

[7] http://alias-i.com/lingpipe/

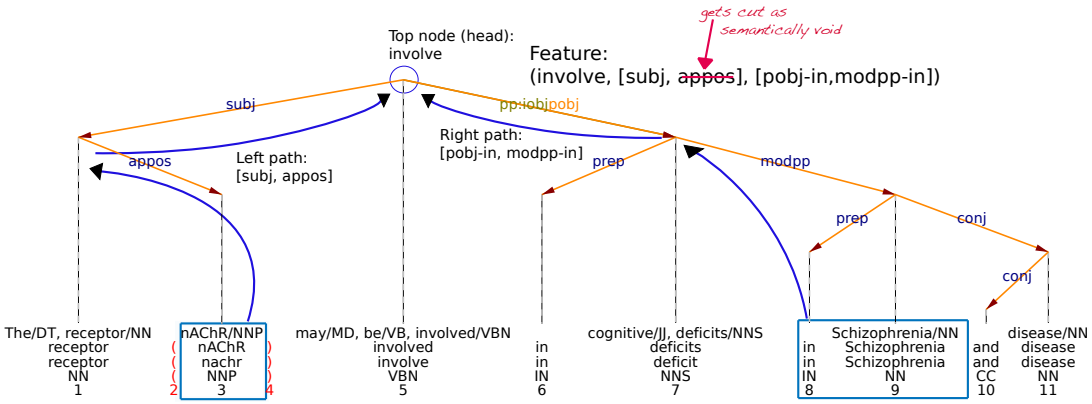


Figure 4: Simplified internal syntactic representation of the sentence "The neuronal nicotinic acetylcholine receptor alpha7 (nAChR alpha7) may be involved in cognitive deficits in Schizophrenia and Alzheimer's disease." from training file 15695160. The curved arrows and dark red notes are aimed at illustrating the path feature (see section 2.4).

The tokenizer produces a granular set of tokens, e.g. words that contain a hyphen (such as 'Pop2p-Cdc18p') are split into several tokens, revealing the inner structure of such constructs which would, for example, allow discovery of the interaction mentioned in *"Pop2p-Cdc18p interaction"*. The OntoGene pipeline also includes a step of term annotation and disambiguation [23, 24]. In order to account for possible surface variants, a normalization step is included in the annotation procedure. The pipeline also includes part-of-speech taggers [25], a lemmatizer [26] and a syntactic chunker [27]. A dependency parser [28] is used to detect the syntactic structure of each sentence.

When the pipeline finishes, each input sentence has been annotated with additional information (see Figure 4), which can be briefly summarized as follows: sentences are tokenized and their borders are detected; each sentence and each token have been assigned an ID; each token is lemmatized; tokens

which belong to terms are grouped; each term is assigned a normalized form and a semantic type; tokens and terms are then grouped into chunks; each chunk has a type (NP or VP) and a head token; each sentence is described as a syntactic dependency structure; each dependency occurs between two tokens and has a type. All this information is represented as a set of predicates and stored into the knowledge base of the system, which can then be queried by an application.

The rich annotations generated by the OntoGene pipeline can also be used to generate candidate interactions using a number of different criterias. Each token in the OntoGene annotation framework is assigned a unique identifier. Extracted terms can be related back to their position in text thanks to the unique token identifiers.

We have explored three basic approaches to generate candidate interactions, with the resulting candidates ranked according to simple criterias such as the frequency of the entities involved.

art: All concepts in the same article are combined in all possible pairs.

sent: All concepts in the same sentence are combined in all possible pairs.

syn: Only pairs licensed by our syntax-based approach are considered.

In more detail, for the 'article' and 'sentence' cases we generate all binary combinations of concept identifiers within the selected text unit (whole abstract in the first case, sentence by sentence in the second case). We decided to exclude self-interactions (a combination of a concept with itself) even at the cost of losing some recall. In the version of PharmGKB used by us there

are 647 self-interactions, which amount to 2.56% of the total. However it is not clear to us what exactly these interactions represent and whether they should be there in the first place.

Figure 1 shows the results obtained applying the 'article' and 'sentence' approach to the full set of PharmGKB abstracts. These results are reasonably encouraging as they show that it is possible to reach a recall of slightly more than 60%, which is quite good considering that only abstracts are used.[8] It is reasonable to expect that a significant proportion of the relevant interactions will be reported only in the main text of the articles. If they are not mentioned in the abstract, they will be inaccessible in our experiments.

In order to derive a ranking for the candidate interactions generated by the system, each candidate pair (c_1, c_2) is assigned a score according to the following formula:

$score(c_1, c_2) = (f(c_1) + f(c_2))/f(C)$

where $f(c_1)$ and $f(c_2)$ are the number of times the identifiers c_1 and c_2 are observed in the abstract, while $f(C)$ is the total count of all identifiers in the abstract. Once a score is assigned to each candidate pair, it is possible to filter out the most unlikely candidates, either by setting a threshold value for the score, or by selecting only the N-best candidates. Using one of these methods will result into variable values of Precision, Recall and F-score, depending on the exact value of the score threshold, or N parameter.

[8]These values represent the recall using only the textual information in the title and abstract. For the results presented further on we also add some of the metadata (MeSH terms and chemical substances) which leads to a maximum recall of 69% on the training data set.

We know from our own previous experiments [10] that giving a "boost" to the entities contained in the title can produce a measurable improvement of ranking of the results (measured by the AUC or TAP metrics). We have empirically verified that the best value of such a boost is about 10. This is equivalent to counting the entities in the title ten times, or in other words to treat the title as if it was repeated ten times.

2.4. Syntax-based approach

Approaches towards identification of entity interactions based on their cooccurrence in a given text span are quite common (e.g. [29]). Other approaches apply handcrafted rules, for example regular expressions for surface searches [30], or syntactic patterns on automatically parsed corpora [31, 32]. These approaches typically achieve high precision at the cost of recall. In our previous work, we used manually-constructed syntactic patterns in order to filter candidate protein-protein interactions [33, 10]. This approach was later enhanced with automatic learning of useful syntactic configuration from a training corpus [34, 11]. In the following we describe how such an approach has been adapted to the PharmGKB dataset.

All sentences in the gold standard have been parsed with the Pro3Gres dependency parser [28]. All entities that appear in the same sentence are potentially interacting, so we record the syntactic path that connects them as *candidate path*. If the gold standard contains the information that these two entities really interact, then we mark the path that connects them as a *relevant path*. The number of relevant paths divided by the number of candidate paths gives us the Maximum-Likelihood probability that a path is relevant:

$p(relevant|candidate\ path) = \frac{f(relevant\ path)}{f(candidate\ path)}$

The most frequent path types in the training set are given in table 1. We can use this probability directly during the application phase: whenever two entities occurring in the same sentence of the application corpus, for example a drug and a disease, have a probability of being relevant above a certain threshold, the system reports the interaction. As syntactic path, we record the dependency labels that connect the two entities, and the topmost word connecting them. A sample path is provided in Figure 4.

Except for the lexeme on top of the path, our features are thus less sparse than the ones of many other approaches. If possible, we use a single feature for the entire path. In the majority of cases, we need to split the path into two halves: from the top-word down to one of the entities as feature 1, and from the top-word down the other entity as feature 2. We use lexical information on transparent words to avoid data sparseness, as follows:

- First, entities occurring inside noun chunks are allowed to replace the head of the chunk.
- Second (if still no relevant path exists), the relations for appositions, conjunctions and hyphens are cut.
- Third (if still no relevant path exists), parts of trees which are headed by a transparent word are cut.

A transparent word [35] is a word that does not substantially affect the meaning of a sentence if it is left out. For example, if *drug A affects groups of patients* then the sentence *drug A affects patients*, which does not contain

Precision	Head	Path1	Path2	TP	Count
13.62%	associate	subj	pobj-with	53	389
17.82%	associate	subj modpp-in	pobj-with	31	174
14.57%	effect	modpp-of	modpp-on	22	151
18.92%	effect	modpp-of	modpp-on modpp-of	21	111
20.65%	association	modpp-of	modpp-with	19	92
6.29%	be	obj modpp-of	subj	19	302
17.82%	metabolize	pobj-by	subj	18	101
29.63%	inhibit	pobj-by	subj	16	54
35.71%	associate	subj modpp-in	pobj-with modpp-of	15	42
23.81%	cause	subj modpp-in	obj	15	63
5.02%	be	subj	obj modpp-of	15	299
100.00%	analyze	subj modpp-in	pobj-in modpart pobj-with	14	14

Table 1: Most frequent true positive path types in the training set

the transparent word *group*, has a very similar meaning. We have learnt transparent words using a machine learning approach: words that occur particularly often inside paths are regarded as transparent [34].

The syntactic relation approach (**syn**) uses three additional factors to calculate a score. First, the frequency of the entities in the document, as the most relevant entities in the given document are typically mentioned several times. Reporting interactions based on the frequencies of entities leads to a very high baseline in protein-protein interaction [11]. Second, the probability of the entity types to enter interactions is used. For example, the probability that a drug and a disease in the same sentence have an interaction is relatively high (about 12%), while the probability that two drugs appearing in the same sentence interact is low (about 1%). Third, we use a simple zoning factor: the title is given ten times the weight of the rest of the text.

In order to assess the impact of the syntactic module on its own, we use a version that has fewer backoffs and parameters than the version that

has been used and optimized for protein-protein interaction [11]. A score is assigned to every candidate interaction according to the following formula:

$$pscore(c_1, c_2) = p(relevant|candidate\ path) * f(c_1) * f(c_2) *$$
$$* p(relevant|entitytypes) * zoningfactor$$

The syntactic approach in its current version only has two backoffs: it splits the path into a left and right half, and transparent words are filtered. We have reduced the number of backoffs in order to keep the effects of the syntax separable, in order to not replicate the methods used for the art and sent approaches.

The syntactic module on its own generally achieves higher precision than the other approaches, but low recall. The two backoffs reduce sparseness, leading to better recall, but at the cost of a considerable drop in precision.

The syntactic approach is harmed by data sparseness, by the fact that not necessarily all of the PhamGKB relationships can be assumed to be manually validated, and by the fact that many interactions are expressed very indirectly. For example, in many cases the two interacting entities do not occur in the same sentence. From the 7658 binary interactions in the gold standard that remain after filtering the 75 evaluation documents and all documents that have more than 20 interactions (see section 2.5), the syntactic training module learns 7229 path tokens where the two entities are found in the same syntactic span. These path tokens fall into 5285 types. Only 889 types (17%) occur more than once.

The sparseness is partly due to term recognition (both entities need to be recognized and grounded correctly) and partly due to interactions across sentence boundaries. The most frequent true positive types are given in table

Precision	Head	Path1	Path2	TP	Count
100.00%	analyze	subj modpp-in	pobj-in modpart pobj-with	14	14
100.00%	investigate	subj modpp-of	sentobj obj modpp-with modpp-of	12	12
100.00%	effect	bridge modpp-of	modpp-on modpp-of	6	6
100.00%	determine	bridge	subj nchunk modpp-for modpp-of	5	5
100.00%	involve	subj	pobj-in modpp-in	4	4
90.00%	disease	nchunk	chunk(genes)	9	10
88.89%	explain	subj	pobj-in	8	9
83.33%	determine	bridge	sentobj subj	5	6
83.33%	catalys	subj	bridge obj	5	6
83.33%	cancer	modpp-in	chunk(risk)	5	6
80.00%	effect	modpp-of	bridge modpp-on modpp-of	4	5
66.67%	metabolise	subj	bridge	4	6
66.67%	measure	sentobj subj modpp-of	bridge	4	6
66.67%	find	obj modpp-between	obj2 modpp-with	4	6
66.67%	determine	subj modpp-in	obj modpp-in modpp-to	4	6
66.67%	correlate	pobj-in	subj	4	6
66.67%	be	pobj-in	obj modpp-of modpp-between	4	6
60.00%	investigate	bridge modpp-of	obj modpp-of	6	10

Table 2: Syntactic paths with high probability of expressing an interaction.

1.

The counts are sorted by inverse frequency. The most frequent path type has 53 instances. Path1 is the half from the top word (Head) of the path to the first entity. Path2 is the half to the second entity. The last column lists how often the path occurs in the entire training corpus, irrespective of whether it expresses relevant interactions or not. The ratio between the last two columns, i.e. the probability $p(relevant|candidate\ path)$, which is the main factor in the syntactic feature, is given in the first column. We can see, for example, that the verb *be* is generally unlikely to head a relevant path, while *cause*, *association*, *associate*, and *analyze* have much higher probabilities. Short and easily interpretable paths such as the first one of table 1 (“*X*

associates with Y") only have relatively low chances of expressing relevant interactions, which indicates that naive implementations of the syntactic feature would have low precision. The very specific and long path in the last row always expresses a relevant interaction. There are 15 paths occurring more than 3 times which are 100% relevant.

On the backoff level, where only one half of the path is recorded, sparseness is a bit less serious. 14558 half-path tokens fall into 5904 types, 2524 of them occur more than once. But a verbal frame that is composed of two separate halves often predicts incorrect complete paths.

Another possible benefit of the syntactic approach is that it detects the lexemes appearing at the top of the path (column 'Head' in the tables), which can be used as keywords for other approaches and may also help to distinguish interaction classes. All paths that are not cases of self-reference and are relevant with at least 60% are given in table 2. Except for *be* in a very specific configuration, all Head words in table 2 are good keyword candidates.

Figure 4 portrays a gold standard interaction which corresponds to the fifth row in table 2. The gene-disease interaction between *'nAChR'* and *'Schizophrenia'* (and also *'Alzheimer's disease'*) is expressed in this sentence. Path1 leads via apposition and subject relation to the verb 'involve'. The apposition relation is semantically void and thus gets cut. Path2 is up from 'Schizophrenia' via the relations modpp-in and pobj-in to 'involve', which is suggested as the head because the paths meet here.

Meth.	Docs	TP	FP	FN	AUC iP/R	n
syn	185	40	145	533	0.106	1
syn	185	55	241	518	0.131	2
syn	185	59	316	514	0.135	3
syn	185	66	410	507	0.140	5
syn	185	67	502	506	0.141	10
syn	185	75	555	498	0.142	all
sent	478	181	297	1575	0.235	1
sent	478	266	683	1490	0.285	2
sent	478	324	1090	1432	0.311	3
sent	478	385	1935	1371	0.328	5
sent	478	460	4023	1296	0.342	10
sent	478	652	30025	1104	0.353	all
art	478	194	284	1570	0.246	1
art	478	292	660	1472	0.301	2
art	478	349	1076	1415	0.327	3
art	478	428	1923	1336	0.348	5
art	478	542	4061	1222	0.371	10
art	478	884	63104	880	0.391	all

Table 3: Results on the 10% evaluation data set, containing a total of 485 documents. The first column gives the approach used (see section 2.3). The second column reports the number of documents with a least one response hit – note that the syntactic approach has far more zero hits (therefore fewer evaluated documents, but also the article approach cannot find any relation in 7 articles). The third to the fifth columns give true positives (TP), false positives (FP) and false negatives (FN). The sixth column contains the macro averaged AUC iP/R. The seventh column contains the cut-off value n used by the BioCreative evaluation tool as a threshold on the number of response hits when computing these results. In rows with $n = all$ no threshold was applied.

2.5. Evaluation Results

For a systematic evaluation using the supervised methods described before, we split the corpus into 90% training data (4540 articles) and 10% test data (505 articles). Because the relation types are distributed unevenly over all documents, we tried to ensure an approximately similar distribution of different relation types in the two data sets.

Table 3 compares the performance of the three basic approaches as computed by the BioCreative evaluation tool with increasing cut-off thresholds thus allowing more and more noise to appear. Note that this tool ignores gold standard annotations for documents where no response hits are generated by the evaluated system. Therefore the results for the syntactic approach rely on a subset of 185 documents. The syntactic approach (**syn**) is hampered severely by the low recall and small number of true positives. The purely frequency based approaches are almost equivalent for threshold 1, however using the full context of an abstract (**art**) generally gives better ranking (AUC iP/R) and recall than using only concept pairs appearing in the same sentence (**sent**). Figure 5 visualizes the same findings as performance curves in terms of precision, recall and F-Score. The high impact of recall on AUC iP/R is obvious in these plots. In Figure 6 we report the performance of the same approaches as above but using the TAP-k metric. As discussed in section 2.2, this is closer to the perspective of a human curator/inspector who will stop using the results of a retrieval system when too many false positives appear.

2.6. Evaluation on a restricted gold standard

In the experiments described so far we have assumed that we could use the interaction dataset provided by PharmGKB as a reliable gold standard. There are however some limitations in this assumption. Although in general PharmGKB provides a high standard of curation, the maintainers of the database do not claim that all the entity pairs that they provide necessarily correspond to an interaction explicitly stated in the original document. Some of the pairs might correspond to a broader type of relationship which is inferred by the curator due to co-occurrence of the entities in a given text span. This limitation could call into question to some extent the validity of the three approaches presented above.

We believe that we can in any case consider these results as scientifically significant under the assumption that we are simply trying to simulate the decisions taken by the PharmGKB curators, rather than trying to capture interactions explicitly stated in the original documents

As a way to better verify the quality of our text mining technologies, we have performed additional experiments using only a small subset of articles where interactions have been explicitly validated. In collaboration with PharmGKB we conducted a separate experiment to test the usefulness of our text mining technologies and curation interface for a rather simple revalidation experiment which is described in detail in [36]. This experiment produced a set of 125 abstract where all interactions have been reliably curated by PharmGKB domain experts.

At the time of the experiments described in this section, manually curated interactions were available for only 75 of those articles. These 75 documents

were used as a test corpus, while the rest of PharmGKB was used for training (excluding however documents which contain more than 20 interactions).

Evaluation results are given in table 4. The method **syn** is identical to the one in table 3, but using the 75 document data evaluation set, and the corresponding training set. Results are clearly better than in table 3, which indicates that the manually verified documents are probably a better gold standard. The method **syn+cooc** includes a sentence coocurrence score, thus obtaining a combination between the **syn** and **sent** method. The method **syn+cooc2** extends the sentence coocurrence score to including the neighboring sentence. The increase in recall indicates that context of more than one sentence is often necessary. The method **syn+cooc2w** weighs the sentence coocurrence score by distance, giving higher scores to entities that appear closer. The method **syn+cooc2wf** is identical but does not use a score threshold, thus returning all results, which increases recall and reduces precision. It aims to give an upper bound on recall. The method **syn+cooc2wb** is identical but uses a relatively high score threshold aiming for a balanced precision/recall output.

These results suggest that syntactic approaches for this particular domain and task need to be combined with other approaches, in our example here shallow co-occurrence, to achieve reasonable recall. This is probably due to the fact that several relations in this domain are expressed very indirectly or involving several sentences. Besides there is considerable data spareness that hinders the effectiveness of our methodology. Advantages of syntactic approaches are that they can achieve good precision and deliver evidence sentences which can be presented to a curator. These tentative conclusions

are of course restricted by the very small amount of documents that were available as test data, and by the fact that the approach was trained on the original PharmGKB resource.

A high-quality interaction resource which can be used as gold standard for a shared task, such as the manually verified documents used for the evaluation described here, can be created from the original PharmGKB data at moderate cost, by using text mining tools and manual filtering, as described in the following section.

Meth.	Docs	TP	FP	FN	AUC iP/R	n	P	R
syn	43	36	149	116	0.215	all	0.307	0.286
syn+cooc	73	116	1044	151	0.277	all	0.143	0.477
syn+cooc2	72	158	2337	106	0.279	all	0.094	0.616
syn+cooc2w	72	165	2685	99	0.286	all	0.091	0.650
syn+cooc2wf	72	23	49	241	0.103	1	0.319	0.103
syn+cooc2wf	72	37	107	227	0.154	2	0.257	0.170
syn+cooc2wf	72	45	171	219	0.175	3	0.208	0.205
syn+cooc2wf	72	67	293	197	0.215	5	0.186	0.312
syn+cooc2wf	72	101	611	163	0.257	10	0.143	0.444
syn+cooc2wf	72	167	3783	97	0.286	all	0.073	0.661
syn+cooc2wb	53	47	180	147	0.220	all	0.270	0.281

Table 4: Results on the 75 manually annotated documents. The first column gives the approach used. The second column reports the number of documents with a least one response hit. The third to the fifth columns give true positives (TP), false positives (FP) and false negatives (FN). The sixth column contains the macro averaged AUC iP/R. The seventh column contains the cut-off value n used by the BioCreative evaluation tool as a threshold on the number of response hits when computing these results. In rows with $n = all$ no threshold was applied. The eighth column reports macro precision, the ninth macro recall.

3. Discussion

Advanced text mining techniques are now reaching a level of maturity that makes them increasingly relevant for the process of curation of biomedical literature. However, the development of effective tools for assisted curation cannot simply be based on accurate text mining, but needs to take into account fundamental Human-Computer Interaction (HCI) research, and requires an understanding of the biological issues that drive the work of the curators. As [37] puts it: "*[...] accurately and comprehensively pulling desired information from text is just the beginning of deploying a text mining system as a database curation tool.*" In this section we discuss previous results on the deployment of text mining systems in the process of biomedical literature curation, and then introduce our ODIN curation system, specifically adapted to the PharmGKB database.

[38, 39] use a manually annotated corpus (gold standard) to simulate an assisted curation environment, where the curators are given either gold standard data or the output of an (imperfect) NLP pipeline. They show that a perfect assisted curation environment would improve the speed of curation by about 33%. Another interesting result is that the curators, although in general considering the results from the NLP tool as helpful, clearly preferred a high recall setting to one chosen to optimize precision or F-score, because it is much easier and less time-consuming to reject incorrect suggestions (false positives) than to add new information from scratch (false negatives). However, a very low precision (i.e. an excessive number of false positives) is equally negative, as it was observed in the interactive task (IAT) of BioCreative III [16], because it would become tedious for the curators to have to

reject too many incorrect suggestions by the system, which are obviously wrong to the human expert.

[40, 41] presents a system ("PaperBrowser") developed for the curators of FlyBase, a database for drosophila genetics and molecular biology. While the document analysis is based on a conventional NLP pipeline, including the dependency parser RASP [42], the curator's interface has been developed in strict collaboration with the end-users. A thorough evaluation is presented, comparing the results of two curators on identical papers in two different experimental conditions: with the full functionalities of the system ("experimental condition") and with a reduced interface corresponding to their traditional analysis approach ("control condition"). Using a set of different metrics, the authors show that the experimental conditions provide the curators with a visible benefit in terms of navigation efficiency and navigation utility.

[37] discuss how well the performance of a text mining system (in their case tailored to identify mentions of protein mutations), when evaluated with conventional techniques, translates into real utility of the system for a curation task. In particular, they compare an 'intrinsic' evaluation scenario (based on a manually curated gold standard, developed specifically for the task), and an 'extrinsic' scenario, where the output of the system is compared against the entries in the database. They find that high performance on gold standard data does not necessarily translate into high performance for database annotation, pointing to the necessity of adopting novel evaluation techniques in order to assess the real utility of text mining tools for the curation effort. They conclude with the suggestion that the way forward

might be the incorporation of automated techniques into a manual annotation process, or alternatively, 'smart' tools for the deposition of annotations could be used to enforce quality criteria even before a curation takes place, i.e. moving the burden increasingly on the authors of the research.

Textpresso is a well-known text mining system which is characterized by the usage of ontological categories of biological concepts [43, 44], as well as by processing full papers. The system functions as a web service where the researchers/curators can submit a query, either keyword-based or category-based (combinations are allowed), can restrict the search to specific zones of the documents (e.g. abstract, title, body, etc.), and can require the keywords to appear all in the same sentence if desired. The category-based search is semantic in nature, because the categories are based on the meaning of the entries and encompass all the known linguistic realizations of those categories (terms). For example, one source of categories (and corresponding terms) is the Gene Ontology (GO). An application of Textpresso for PharmGKB (Pharmspresso) is also available. Curatorial work done with the assistance of Textpresso was shown to be much more efficient than when done by human readers alone. Efficiency was shown to increase dramatically (up to 39-fold in the best case). They state that: "*For biologists, an automated system with high recall and even moderate precision [...] confers a great advantage over skimming text by eye*" [43].

As part of our own research in this area we developed a curation system called "OntoGene Document INspector" (ODIN [45]) which interfaces with our text mining pipeline. We have used a version of ODIN for our participation to the 'interactive curation' task (IAT) of the BioCreative III evaluation

[16]. This was an informal task without a quantitative evaluation of the participating systems. However, the curators who used the system commented extremely positively on its usability for a practical curation task.

More recently, we have created a version of ODIN which allows inspection of abstracts automatically annotated with PharmGKB entities (the annotation is performed using the Ontogene pipeline).[9] Users can access either preprocessed documents, or enter any PubMed identifier and have the corresponding abstract processed "on the fly". For the documents already in PharmGKB it is also possible to inspect the gold standard and compare the results of the system against the gold standard. The curator can inspect all entities annotated by the system, and easily modify them if needed (removing false positives with a simple click, or adding missed terms if necessary). The modified documents can be sent back for reprocessing if desired, obtaining therefore modified candidate interactions. The user can also inspect the set of candidate interactions generated by the system, and act upon them just as on entities, i.e. confirm those which are correct, remove those which are incorrect. Candidate interactions will be presented in a ranked order according to the score which has been assigned to them by the text mining system, therefore the curator can choose to work with only a small set of highly ranked candidates, ignoring all the rest.

ODIN, which is based on a client-server architecture, maintains a log of the interaction with the curator, which could be used for later revision by a supervisor or for reversing some specific annotation decisions. At the end

[9] `http://www.ontogene.org/pharmgkb/`

of a session the modified annotations are sent back to the server, together with the log, for permanent storage, and can be accessed again at the next session, which could take place on a different remote client. Additionally, the curator can choose to export the annotations to a local file in a simplified format (e.g. comma-separated values).[10]

Fully automated extraction of information from the literature is currently unrealistic, but text mining tools are already sufficiently reliable to provide hints to the curators, and have been shown to speed up their activities: *"Although the outputs produced by large-scale IE systems are not yet suitable for producing factual databases for direct use by biomedical researchers, the current level of performance provides two important facilities to the research community. First, the results of these efforts can be used to significantly increase the efficiency of manual curation efforts. Each extracted assertion is tied to a specific text, which can be used to direct the attention of manual curators both to relevant documents and to specific relevant passages within a document."* [46]

4. Conclusion

In this paper we have discussed the possible usage of the PharmGKB as a reference dataset for a relation mining task analogous to the protein-protein interaction task of the 2009 BioCreative competitive evaluation. This might allow the establishment of a relation mining task involving entities such as drugs, diseases and genes.

[10]This functionality is not available in the demo version.

We have shown how to apply existing tools to score the results and provide reliable metrics, including not only the traditional Precision, Recall and F-score but also the increasingly important measures of ranking quality, such as "AUC iP/R" or "TAP-k".

We have presented our own approach towards the mining of pharmacogenomics relationships and scored it against the PharmGKB dataset. Our experiments show that this task is feasible, and our results might offer a useful baseline for further developments in this area. Finally, we have presented an implementation of our assisted curation environment (ODIN) specifically adapted to the PharmGKB dataset.

5. Acknowledgements

This research is partially funded by the Swiss National Science Foundation (grant 100014-118396/1). Additional support is provided by Novartis Pharma AG, NITAS, Text Mining Services, CH-4002, Basel, Switzerland. The authors wish to thank Kevin Cohen and Larry Hunter for precious comments and suggestions.

[1] W. A. Baumgartner, K. B. Cohen, L. M. Fox, G. Acquaah-Mensah, L. Hunter, Manual curation is not sufficient for annotation of genomic databases, Bioinformatics 23 (13) (2007) i41–48. arXiv:http://bioinformatics.oxfordjournals.org/cgi/reprint/23/13/i41.pdf, doi:10.1093/bioinformatics/btm229.
URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/23/13/i41

[2] M. Krallinger, F. Leitner, C. Rodriguez-Penagos, A. Valencia, Overview of the protein-protein interaction annotation extraction task of BioCreative II, Genome Biology 9 (Suppl 2) (2008) S4.

[3] F. Leitner, S. A. Mardis, M. Krallinger, G. Cesareni, L. A. Hirschman, A. Valencia, An overview of biocreative ii.5, IEEE/ACM Transactions on Computational Biology and Bioinformatics 7 (3) (2010) 385–399. doi:10.1109/TCBB.2010.50.

[4] J.-D. Kim, T. Ohta, S. Pyysalo, Y. Kano, J. Tsujii, Overview of bionlp'09 shared task on event extraction, in: Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task, 2009, pp. 1–9.

[5] D. Rebholz-Schuhmann, A. Yepes, C. Li, S. Kafkas, I. Lewin, N. Kang, P. Corbett, D. Milward, E. Buyko, E. Beisswanger, K. Hornbostel, A. Kouznetsov, R. Witte, J. Laurila, C. Baker, C.-J. Kuo, S. Clematide, F. Rinaldi, R. Farkas, G. Mora, K. Hara, L. I. Furlong, M. Rautschka, M. Neves, A. Pascual-Montano, Q. Wei, N. Collier, M. Chowdhury, A. Lavelli, R. Berlanga, R. Morante, V. Van Asch, W. Daelemans, J. Marina, E. van Mulligen, J. Kors, U. Hahn, Assessment of ner solutions against the first and second calbc silver standard corpus, Journal of Biomedical Semantics 2 (Suppl 5) (2011) S11. doi:10.1186/2041-1480-2-S5-S11.
URL http://www.jbiomedsem.com/content/2/S5/S11

[6] H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, R. Apweiler, IntAct: an

open source molecular interaction database, Nucl. Acids Res. 32 (suppl 1) (2004) D452–455. arXiv:http://nar.oxfordjournals.org/cgi/reprint/32/suppl_1/D452.pdf, doi:10.1093/nar/gkh052.
URL http://nar.oxfordjournals.org/cgi/content/abstract/32/suppl_1/D452

[7] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, G. Cesareni, MINT: a Molecular INTeraction database, FEBS Letters 513 (1) (2002) 135–140.

[8] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, M. Tyers, Biogrid: A general repository for interaction datasets, Nucleic Acids Research 34 (2006) D535–9.

[9] J. Kim, T. Ohta, Y. Tateisi, J. Tsujii, GENIA corpus - a semantically annotated corpus for bio-textmining, Bioinformatics 19 (1) (2003) 180–182.
URL http://bioinformatics.oupjournals.org/cgi/content/abstract/19/suppl_1/i180

[10] F. Rinaldi, T. Kappeler, K. Kaljurand, G. Schneider, M. Klenner, S. Clematide, M. Hess, J.-M. von Allmen, P. Parisot, M. Romacker, T. Vachon, OntoGene in BioCreative II, Genome Biology 9 (Suppl 2) (2008) S13. doi:10.1186/gb-2008-9-s2-s13.
URL http://genomebiology.com/2008/9/S2/S13

[11] F. Rinaldi, G. Schneider, K. Kaljurand, S. Clematide, T. Vachon, M. Romacker, OntoGene in BioCreative II.5, IEEE/ACM Transactions on Computational Biology and Bioinformatics 7 (3) (2010) 472–480. doi:10.1109/TCBB.2010.50.

[12] T. Kappeler, S. Clematide, K. Kaljurand, G. Schneider, F. Rinaldi, Towards automatic detection of experimental methods from biomedical literature, in: T. Salakoski, D. Rebholz-Schuhmann, S. Pyysalo (Eds.), Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008), Turku, Finland, Turku Centre for Computer Science (TUCS), 2008, pp. 61–68.

[13] G. Schneider, S. Clematide, F. Rinaldi, Detection of interaction articles and experimental methods in biomedical literature, BMC Bioinformatics 12 (Suppl 8) (2011) S13. doi:10.1186/1471-2105-12-S8-S13.
URL http://www.biomedcentral.com/1471-2105/12/S8/S13

[14] Z. Lu, H.-Y. Kao, C.-H. Wei, M. Huang, J. Liu, C.-J. Kuo, C.-N. Hsu, R. Tsai, H.-J. Dai, N. Okazaki, H.-C. Cho, M. Gerner, I. Solt, S. Agarwal, F. Liu, D. Vishnyakova, P. Ruch, M. Romacker, F. Rinaldi, S. Bhattacharya, P. Srinivasan, H. Liu, M. Torii, S. Matos, D. Campos, K. Verspoor, K. Livingston, W. Wilbur, The gene normalization task in biocreative iii, BMC Bioinformatics 12 (Suppl 8) (2011) S2. doi:10.1186/1471-2105-12-S8-S2.
URL http://www.biomedcentral.com/1471-2105/12/S8/S2

[15] M. Krallinger, M. Vazquez, F. Leitner, D. Salgado, A. Chatr-aryamontri, A. Winter, L. Perfetto, L. Briganti, L. Licata, M. Iannuccelli, L. Castagnoli, G. Cesareni, M. Tyers, G. Schneider, F. Rinaldi, R. Leaman, G. Gonzalez, S. Matos, S. Kim, W. Wilbur, L. Rocha, H. Shatkay, A. Tendulkar, S. Agarwal, F. Liu, X. Wang, R. Rak, K. Noto, C. Elkan, Z. Lu, R. Dogan, J.-F. Fontaine, M. Andrade-Navarro, A. Valencia, The protein-protein interaction tasks of biocreative iii: classification/ranking of articles and linking bio-ontology concepts to full text, BMC Bioinformatics 12 (Suppl 8) (2011) S3. doi:10.1186/1471-2105-12-S8-S3.
URL http://www.biomedcentral.com/1471-2105/12/S8/S3

[16] C. Arighi, P. Roberts, S. Agarwal, S. Bhattacharya, G. Cesareni, A. Chatr-aryamontri, S. Clematide, P. Gaudet, M. Giglio, I. Harrow, E. Huala, M. Krallinger, U. Leser, D. Li, F. Liu, Z. Lu, L. Maltais, N. Okazaki, L. Perfetto, F. Rinaldi, R. Saetre, D. Salgado, P. Srinivasan, P. Thomas, L. Toldo, L. Hirschman, C. Wu, Biocreative iii interactive task: an overview, BMC Bioinformatics 12 (Suppl 8) (2011) S4. doi:10.1186/1471-2105-12-S8-S4.
URL http://www.biomedcentral.com/1471-2105/12/S8/S4

[17] J. Davis, M. Goadrich, The relationship between precision-recall and roc curves, in: ICML '06: Proceedings of the 23rd international conference

on Machine learning, ACM, New York, NY, USA, 2006, pp. 233–240. doi:http://doi.acm.org/10.1145/1143844.1143874.

[18] T. Klein, J. Chang, M. Cho, K. Easton, R. Fergerson, M. Hewett, Z. Lin, Y. Liu, S. Liu, D. Oliver, D. Rubin, F. Shafa, J. Stuart, R. Altman, Integrating genotype and phenotype information: An overview of the PharmGKB project, The Pharmacogenomics Journal 1 (2001) 167–170.

[19] K. Sangkuhl, D. S. Berlin, R. B. Altman, T. E. Klein, PharmGKB: Understanding the effects of individual genetic variants, Drug Metabolism Reviews 40 (4) (2008) 539–551, pMID: 18949600. arXiv:http://informahealthcare.com/doi/pdf/10.1080/03602530802413338, doi:10.1080/03602530802413338.
URL http://informahealthcare.com/doi/abs/10.1080/03602530802413338

[20] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[21] H. D. Carroll, M. G. Kann, S. L. Sheetlin, J. L. Spouge, Threshold average precision (tap-k): a measure of retrieval designed for bioinformatics, Bioinformatics 26 (14) (2010) 1708–1713. doi:10.1093/bioinformatics/btq270.

[22] F. Rinaldi, G. Schneider, K. Kaljurand, M. Hess, M. Romacker, An Environment for Relation Mining over Richly Annotated Corpora: the case of GENIA, BMC Bioinformatics 7 (Suppl 3) (2006) S3. doi:10.1186/1471-2105-7-S3-S3.
URL http://www.biomedcentral.com/1471-2105/7/S3/S3

[23] F. Rinaldi, K. Kaljurand, R. Saetre, Terminological resources for text mining over biomedical scientific literature, Journal of Artificial Intelligence in Medicine 52 (2) (2011) 107–114.

[24] K. Kaljurand, F. Rinaldi, T. Kappeler, G. Schneider, Using existing biomedical resources to detect and ground terms in biomedical literature, in: Proceedings of the 12th Conference on Artificial Intelligence in Medicine (AIME09), 2009, pp. 225–234.

[25] Y. Tsuruoka, Y. Tateishi, J.-D. Kim, T. Ohta, J. McNaught, S. Ananiadou, J. Tsujii, Developing a robust part-of-speech tagger for biomedical text, in: Advances in Informatics - 10th Panhellenic Conference on Informatics, LNCS 3746, 2005, pp. 382–392.

[26] G. Minnen, J. Carroll, D. Pearce, Applied morphological processing of English, Natural Language Engineering 7 (3) (2001) 207–223.

[27] A. Mikheev, S. Finch, A workbench for finding structure in texts, in: Proceedings of the Fifth Conference on Applied Natural Language Processing, Association for Computational Linguistics, Washington, DC, USA, 1997, pp. 372–379. doi:10.3115/974557.974611.
URL http://www.aclweb.org/anthology/A97-1054

[28] G. Schneider, Hybrid long-distance functional dependency parsing, Doctoral Thesis, Institute of Computational Linguistics, University of Zurich (2008).

[29] D. Rebholz-Schuhmann, H.Kirsch, M. Arregui, S. Gaudan, M. Riethoven, P.Stoehr, EBIMed – text crunching to gather facts for proteins from Medline, Bioinformatics 23(2) (2006) e237 – e244.

[30] C. Giuliano, A. Lavelli, L. Romano, Exploiting shallow linguistic information for relation extraction from biomedical literature, in: Proceedings of EACL 2006, 2006, pp. 401–408.

[31] F. Rinaldi, G. Schneider, K. Kaljurand, M. Hess, M. Romacker, An environment for relation mining over richly annotated corpora: the case of GENIA, BMC Bioinformatics 7(Suppl 3):S3.

[32] K. Fundel, R. Küffner, R. Zimmer, RelEx – relation extraction extraction using dependency parse trees, Bioinformatics 23(3) (2007) 365–371.

[33] F. Rinaldi, G. Schneider, K. Kaljurand, M. Hess, C. Andronis, O. Konstandi, A. Persidis, Mining of Functional Relations between Genes and Proteins over Biomedical Scientific Literature using a Deep-Linguistic Approach, Journal of Artificial Intelligence in Medicine 39 (2007) 127–136. doi:10.1016/j.artmed.2006.08.005.

[34] G. Schneider, K. Kaljurand, F. Rinaldi, Detecting Protein/Protein Interactions using a parser and linguistic resources, in: CICLing 2009, 10th International Conference on Intelligent Text Processing and Computational Linguistics, Springer LNC 5449, Mexico City, Mexico, 2009, pp. 406–417.

[35] A. Meyers, Annotation guidelines for nombank – noun argument structure for propbank (undated).

[36] F. Rinaldi, S. Clematide, Y. Garten, M. Whirl-Carrillo, L. Gong, J. M. Hebert, K. Sangkuhl, C. F. Thorn, T. E. Klein, R. B. Altman, Using ODIN for a PharmGKB re-validation experiment, Database: The Journal of Biological Databases and Curationdoi:10.1093/database/bas021.

[37] J. G. Caporaso, N. Deshpande, J. L. Fink, E. Bourne, K. B. Cohen, L. Hunter, Intrinsic evaluation of text mining tools may not predict performance on realistic tasks., in: Pacific Symposium on Biocomputing 13, 2008, pp. 640–651.
URL http://view.ncbi.nlm.nih.gov/pubmed/18229722

[38] B. Alex, C. Grover, B. Haddow, M. Kabadjov, E. Klein, M. Matthews, S. Roebuck, R. Tobin, X. Wang, Assisted curation: Does text mining really help, in: R. B. Altman, A. K. Dunker, L. Hunter, T. Murray, T. E. Klein (Eds.), BIOCOMPUTING 2008. Proceedings of the Pacific Symposium on Biocomputing, Kohala Coast, Hawaii, USA, 2008, pp. 556–567.
URL http://psb.stanford.edu/psb-online/proceedings/psb08/alex.pdf

[39] B. Alex, C. Grover, B. Haddow, M. Kabadjov, E. Klein, M. Matthews, R. Tobin, X. Wang, Automating curation using a natural language processing pipeline, Genome Biology 9 (Suppl 2) (2008) S10.

[40] N. Karamanis, R. Seal, I. Lewin, P. McQuilton, A. Vlachos, C. Gasperin, R. Drysdale, T. Briscoe, Natural language processing in aid of flybase curators, BMC Bioinformatics 9 (1) (2008) 193. doi:10.1186/1471-2105-9-193.

[41] N. Karamanis, I. Lewin, R. Seal, R. A. Drysdale, E. J. Briscoe, Integrating natural language processing with flybase curation, in: Pacific Symposium on Biocomputing, 2007, pp. 245–256.

[42] T. Briscoe, J. Carroll, R. Watson, The second release of the RASP system, in: Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, Association for Computational Linguistics, Sydney, Australia, 2006, pp. 77–80.
URL http://www.aclweb.org/anthology/P/P06/P06-4020

[43] H.-M. Müller, E. E. Kenny, P. W. Sternberg, Textpresso: An ontology-based information retrieval and extraction system for biological literature, PLoS Biol 2 (11) (2004) e309. doi:10.1371/journal.pbio.0020309.
URL http://dx.doi.org/10.1371\%2Fjournal.pbio.0020309

[44] H. Müller, A. Rangarajan, T. Teal, P. Sternberg, Textpresso for neuroscience: searching the full text of thousands of neuroscience research papers., Neuroinformatics 6 (3) (2008) 195–20.

[45] F. Rinaldi, S. Clematide, G. Schneider, M. Romacker, T. Vachon, ODIN: An advanced interface for the curation of biomedical literature, in: Biocuration 2010, the Conference of the International Society for Biocuration and the 4th International Biocuration Conference., 2010, p. 61.

[46] L. Hunter, Z. Lu, J. Firby, W. Baumgartner, H. Johnson, P. Ogren, K. B. Cohen, OpenDMAP: An open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression, BMC Bioinformatics 9 (1) (2008) 78. doi:10.1186/1471-2105-9-78.
URL http://www.biomedcentral.com/1471-2105/9/78

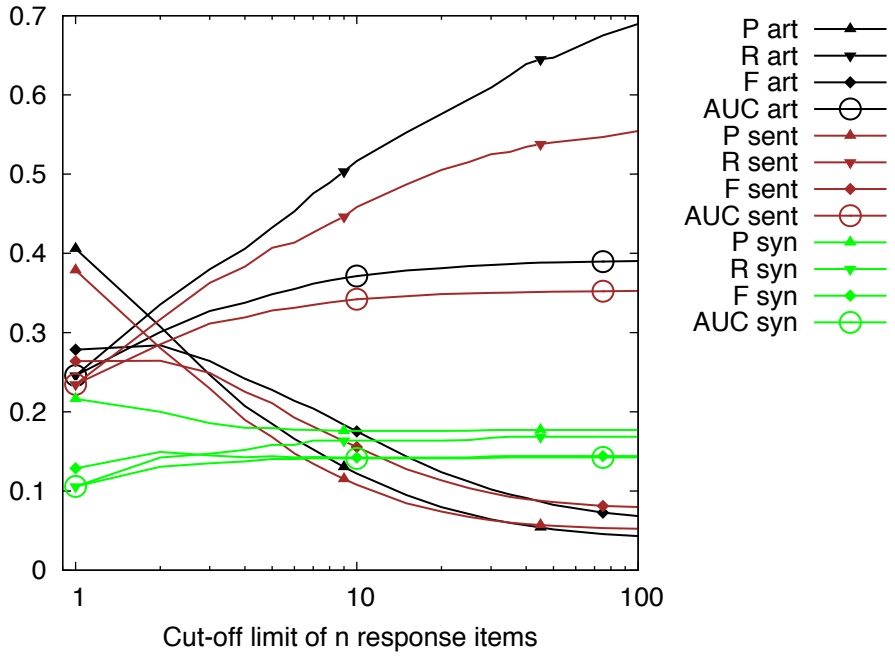


Figure 5: Results computed using the BioCreative evaluation tool for our different relation mining approaches on the 10% evaluation data set. The horizontal axis shows the cut-off value limiting the number of hits that are evaluated by the tool. The vertical axis shows macro averaged results of precision (P), recall (R), F-score (F) and AUC iP/R for our different approaches. Note that these results were computed by ignoring documents without hits in the system responses (this is the default setting for the BioCreative evaluations). See table 3 for the number of documents that produce hits.

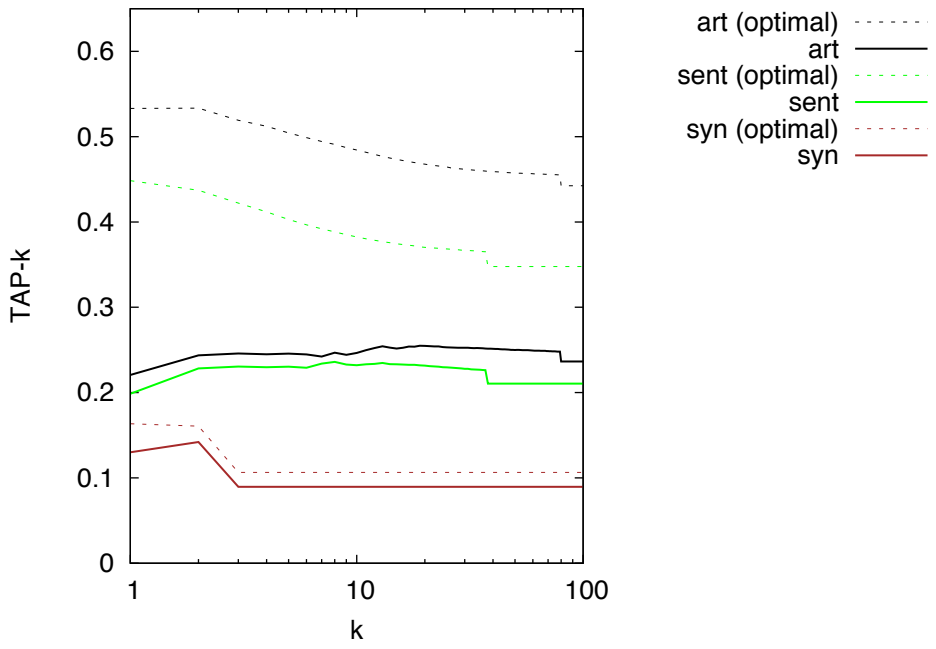


Figure 6: The TAP-k values for our three approaches on the 10% evaluation data set. The horizontal axis shows the k threshold. The vertical axis shows the resulting TAP for a given k. Note that the flat segment is due to the padding of the result list with dummy results if too few results are reported for a query to reach k. This affects especially the syntactic and sentential approaches that deliver far less results than the article approach. The dotted lines show the TAP-k values which could be achieved if all true positive hits of the system would be ranked optimally as hits with the highest confidence.

Figure 7: Example of interaction with the ODIN system. Candidate interactions are listed in the right-hand-side panel. When the user selects one of those interactions, the terms which contribute to its identification are highlighted in the abstract.